EXHIBIT A
U.S. Serial No. 09/981,124
Filed: October 17, 2001
Applicants: Allan Green, et al.

BLAST HELP MANUAL

DESCRIPTION

This document describes the WWW BLAST interface.

BLAST (Basic Local Alignment Search Tool) is the heuristic search algorithm employed by the programs blastp, blastn, blastx, tblastn, and tblastx: these programs ascribe significance to their findings using the statistical methods of Karlin and Altschul (1990, 1993) with a few enhancements. The BLAST programs were tailored for sequence similarity searching -- for example to identify homologs to a query sequence. The programs are not generally useful for motif-style searching. For a discussion of basic issues in similarity searching of sequence databases, see Altschul et al. (1994).

The five BLAST programs described here perform the following tasks:

blastp compares an amino acid query sequence against a protein sequence database;

blastn compares a nucleotide query sequence against a nucleotide sequence database;

blastx compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database;

tblastn compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).

tblastx compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

BLAST Search parameters

HISTOGRAM

Display a histogram of scores for each search; default is yes. (See parameter H in

DESCRIPTIONS

Restricts the number of short descriptions of matching sequences reported to the r default limit is 100 descriptions. (See parameter V in the manual page). See also I

Applicants: Allan Green et al.
Serial No.: 09/981,124
Filed: October 17, 2001
Exhibit A

CUTOFF.

ALIGNMENTS

Restricts database sequences to the number specified for which high-scoring segn reported; the default limit is 50. If more database sequences than this happen to sa significance threshold for reporting (see EXPECT and CUTOFF below), only the greatest statistical significance are reported. (See parameter B in the BLAST Man

EXPECT

The statistical significance threshold for reporting matches against database seque is 10, such that 10 matches are expected to be found merely by chance, according model of Karlin and Altschul (1990). If the statistical significance ascribed to a m EXPECT threshold, the match will not be reported. Lower EXPECT thresholds a leading to fewer chance matches being reported. Fractional values are acceptable. the BLAST Manual).

CUTOFF

Cutoff score for reporting high-scoring segment pairs. The default value is calcul value (see above). HSPs are reported for a database sequence only if the statistica to them is at least as high as would be ascribed to a lone HSP having a score equa value. Higher CUTOFF values are more stringent, leading to fewer chance match parameter S in the BLAST Manual). Typically, significance thresholds can be mo using EXPECT.

MATRIX

Specify an alternate scoring matrix for BLASTP, BLASTX, TBLASTN and TBL matrix is BLOSUM62 (Henikoff & Henikoff, 1992). The valid alternative choice PAM120, PAM250 and IDENTITY. No alternate scoring matrices are available f specifying the MATRIX directive in BLASTN requests returns an error response.

STRAND

Restrict a TBLASTN search to just the top or bottom strand of the database seque BLASTN, BLASTX or TBLASTX search to just reading frames on the top or bot query sequence.

FILTER

Mask off segments of the query sequence that have low compositional complexity SEG program of Wootton & Federhen (Computers and Chemistry, 1993), or segn short-periodicity internal repeats, as determined by the XNU program of Claverie and Chemistry, 1993), or, for BLASTN, by the DUST program of Tatusov and Li Filtering can eliminate statistically significant but biologically uninteresting repor (e.g., hits against common acidic-, basic- or proline-rich regions), leaving the mor interesting regions of the query sequence available for specific matching against d Low complexity sequence found by a filter program is substituted using the letter sequence (e.g., "NNNNNNNNNNNNN") and the letter "X" in protein sequences "XXXXXXXXX"). Users may turn off filtering by using the "Filter" option on th for the BLAST server" page.

Filtering is only applied to the query sequence (or its translation products), not to Default filtering is DUST for BLASTN, SEG for other programs.

It is not unusual for nothing at all to be masked by SEG, XNU, or both, when app SWISS-PROT, so filtering should not be expected to always yield an effect. Furth sequences are masked in their entirety, indicating that the statistical significance c against the unfiltered query sequence should be suspect.

NCBI-gi

Causes NCBI gi identifiers to be shown in the output, in addition to the accession

SEARCH STRATEGY

The fundamental unit of BLAST algorithm output is the High-scoring Segment Pair (HSP). An HSP consists of two sequence fragments of arbitrary but equal length whose alignment is locally maximal and for which the alignment score meets or exceeds a threshold or cutoff score. A set of HSPs is thus defined by two sequences, a scoring system, and a cutoff score; this set may be empty if the cutoff score is sufficiently high. In the programmatic implementations of the BLAST algorithm described here, each HSP consists of a segment from the query sequence and one from a database sequence. The sensitivity and speed of the programs can be adjusted via the standard BLAST algorithm parameters W, T, and X (Altschul et al., 1990); selectivity of the programs can be adjusted via the cutoff score.

A Maximal-scoring Segment Pair (MSP) is defined by two sequences and a scoring system and is the highest-scoring of all possible segment pairs that can be produced from the two sequences. The statistical methods of Karlin and Altschul (1990, 1993) are applicable to determining the significance of MSP scores in the limit of long sequences, under a random sequence model that assumes independent and identically distributed choices for the residues at each position in the sequences. In the programs described here, Karlin-Altschul statistics have been extrapolated to the task of assessing the significance of HSP scores obtained from comparisons of potentially short, biological sequences.

The approach to similarity searching taken by the BLAST programs is first to look for similar segments (HSPs) between the query sequence and a database sequence, then to evaluate the statistical significance of any matches that were found, and finally to report only those matches that satisfy a user-selectable threshold of significance. Findings of multiple HSPs involving the query sequence and a single database sequence may be treated statistically in a variety of ways. By default the programs use "Sum" statistics (Karlin and Altschul, 1993). As such, the statistical significance ascribed to a set of HSPs may be higher than that ascribed to any individual member of the set. Only when the ascribed significance satisfies the user-selectable threshold (E parameter) will the match be reported to the user.

The task of finding HSPs begins with identifying short words of length W in the query sequence that either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul et al., 1990). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached.

KARLIN-ALTSCHUL STATISTICS

From Karlin and Altschul (1990), the principal equation relating the score of an HSP to its expected frequency of chance occurrence is:

$$E = K N \exp(-\text{Lambda } S)$$

where E is the expected frequency of chance occurrence of an HSP having score S (or one scoring higher); K and Lambda are Karlin-Altschul parameters; N is the product of the query and database sequence lengths, or the size of the search space; and exp is the exponentiation function.

Lambda may be thought of as the expected increase in reliability of an alignment associated with a unit increase in alignment score. Reliability in this case is expressed in units of information, such as bits or nats, with one nat being equivalent to 1/log(2) (roughly 1.44) bits.

The expectation E (range 0 to infinity) calculated for an alignment between the query sequence and a database sequence can be extrapolated to an expectation over the entire database search, by converting the pairwise expectation to a probability (range 0-1) and multiplying the result by the ratio of the entire database size (expressed in residues) to the length of the matching database sequence. In detail:

$$E\_database = (1 - \exp(-E)) \; D / d$$

where D is the size of the database; d is the length of the matching database sequence; and the quantity (1 - exp(-E)) is the probability, P, corresponding to the expectation E for the pairwise sequence comparison. Note that in the limit of infinite E, P approaches 1; and in the limit as E approaches 0, E and P approach equality. Due to inaccuracy in the statistical methods as they are applied in the BLAST programs, whenever E and P are less than about 0.05, the two values can be practically treated as being equal.

In contrast to the random sequence model used by Karlin-Altschul statistics, biological sequences are often short in length -- an HSP may involve a relatively large fraction of the query or database sequence, which reduces the effective size of the 2-dimensional search space defined by the two sequences. To obtain more accurate significance estimates, the BLAST programs compute effective lengths for the query and database sequences that are their real lengths minus the expected length of the HSP, where the expected length for an HSP is computed from its score. In no event is an effective length for the query or database sequence permitted to go below 1. Thus, the effective length of either the query or the database sequence is computed according to the following:

$$Length\_eff = MAX(Length\_real - Lambda \; S / H \; , \; 1)$$

where H is the relative entropy of the target and background residue frequencies (Karlin and Altschul, 1990), one of the

statistics reported by the BLAST programs. H may be thought of as the information expected to be obtained from each pair of aligned residues in a real alignment that distinguishes the alignment from a random one.

SCORING SCHEMES

The default scoring matrix used by blastp, blastx, tblastn, and tblastx is the BLOSUM62 matrix (Henikoff and Henikoff, 1992).

Several PAM (point accepted mutations per 100 residues) amino acid scoring matrices are provided in the BLAST software distribution, including the PAM40, PAM120, and PAM250. While the BLOSUM62 matrix is a good general purpose scoring matrix and is the default matrix used by the BLAST programs, if one is restricted to using only PAM scoring matrices, then the PAM120 is recommended for general protein similarity searches (Altschul, 1991). The pam(1) program can be used to produce PAM matrices of any desired iteration from 2 to 511. Each matrix is most sensitive at finding similarities at its particular PAM distance. For more thorough searches, particularly when the mutational distance between potential homologs is unknown and the significance of their similarity may be only marginal, Altschul (1991, 1992) recommends performing at least three searches, one each with the PAM40, PAM120 and PAM250 matrices.

In blastn, the M parameter sets the reward score for a pair of matching residues; the N parameter sets the penalty score for mismatching residues. M and N must be positive and negative integers, respectively. The relative magnitudes of M and N determines the number of nucleic acid PAMs (point accepted mutations per 100 residues) for which they are most sensitive at finding homologs. Higher ratios of M:N correspond to increasing nucleic acid PAMs (increased divergence). The default values for M and N, respectively 5 and -4, having a ratio of 1.25, correspond to about 47 nucleic acid PAMs, or about 58 amino acid PAMs; an M:N ratio of 1 corresponds to 30 nucleic acid PAMs or 38 amino acid PAMs. At higher than about 40 nucleic acid PAMs, or 50 amino acid PAMs, better sensitivity at detecting similarities between coding regions is expected by performing comparisons at the amino acid level (States et al., 1991), using conceptually translated nucleotide sequences (re: blastx, tblastn, and tblastx).

Independent of the values chosen for M and N, the default wordlength W=11 used by blastn restricts the program to finding sequences that share at least an 11-mer stretch of 100% identity with the query. Under the random sequence model, stretches of 11 consecutive matching residues are unlikely to occur merely by chance even between only moderately diverged homologs. Thus, blastn with its default parameter settings is poorly suited to finding anything but very similar sequences. If better sensitivity is needed, one should use a smaller value for W.

For the blastn program, it may be easy to see how multiply-

ing both M and N by some large number will yield proportionally larger alignment scores with their statistical significance remaining unchanged. This scale-independence of the statistical significance estimates from blastn has its analog in the scoring matrices used by the other BLAST programs: multiplying all elements in a scoring matrix by an arbitrary factor will proportionally alter the alignment scores but will not alter their statistical significance (assuming numerical precision is maintained). From this it should be clear that raw alignment scores are meaningless without specific knowledge of the scoring matrix that was used.

SCORING REQUIREMENTS

Regardless of the scoring scheme employed, two stringent criteria must be met in order to be able to calculate the Karlin-Altschul parameters Lambda and K. First, given the residue composition for the query sequence and the residue composition assumed for the database, the alignment score expected for any randomly selected pair of residues (one from the query sequence and one from the database) must be negative. Second, given the sequence residue compositions and the scoring scheme, a positive score must be possible to achieve. For instance, the match reward score of blastn must have a positive value; and given the assumption made by blastn that the 4 nucleotides A, C, G and T are represented at equal 25% frequencies in the database, a wide range of value combinations for M and N are precluded from use -- namely those combinations where the magnitude of the ratio M:N is greater than or equal to 3.

GENETIC CODES

The parameter C can be set to a positive integer to select the genetic code that will be used by blastx and tblastx to translate the query sequence. The -dbgcode parameter can be used to select an alternate genetic code for translation of the database by the programs tblastn and tblastx. In each case, the default genetic code is the so-called "Standard" or "Universal" genetic code. To obtain a listing of the genetic codes available and their associated numerical identifiers, invoke blastx or tblastx with the command line parameter C=list. Note: the numerical identifiers used here for genetic codes parallel those defined in the NCBI software Toolbox; hence some numerical values will be skipped as genetic codes are updated.

The list of genetic codes available and their associated values for the parameters C and -dbgcode are:

1 Standard or Universal

2 Vertebrate Mitochondrial

3 Yeast Mitochondrial

4 Mold, Protozoan, Coelenterate Mitochondrial and

Mycoplasma/Spiroplasma

5 Invertebrate Mitochondrial

6 Ciliate Macronuclear

9 Echinodermate Mitochondrial

10 Alternative Ciliate Macronuclear

11 Eubacterial

12 Alternative Yeast

13 Ascidian Mitochondrial

14 Flatworm Mitochondrial

P-VALUES, ALIGNMENT SCORES, AND INFORMATION

The Expect and P-values reported for HSPs are dependent on several factors including: the scoring system employed, the residue composition of the query sequence, an assumed residue composition for a typical database sequence, the length of the query sequence, and the total length of the database. HSP scores from different program invocations are appropriate for comparison even if the databases searched are of different lengths, as long as the other factors mentioned here do not vary. For example, alignment scores from searches with the default BLOSUM62 matrix should not be directly compared with scores obtained with the PAM120 matrix; and scores produced using two versions of the same PAM matrix, each created to different scales (see above), can not be meaningfully compared without conversion to the same scale.

Some isolation from the many factors involved in assessing the statistical significance of HSPs can be attained by observing the information content reported (in bits) for the alignments. While the information content of an HSP may change when different scoring systems are used (e.g., with different PAM matrices), the number of bits reported for an HSP will at least be independent of the scale to which the scoring matrix was generated. (In practice, this statement is not quite true, because the alignment scores used by the BLAST programs are integers that lack much precision). In other words, when conveying the statistical significance of an alignment, the alignment score itself is not useful unless the specific scoring matrix that was employed is also provided, but the informativeness of an alignment is a meaningful statistic that can be used to ascribe statistical significance (a P-value) to the match independently of specific knowledge about the scoring matrix.

SAMPLE OUTPUT

The BLAST programs all provide information in roughly the

same format. First comes (A) an introduction to the program; (B) a histogram of expectations (see above) if one was requested; (C) a series of one-line descriptions of matching database sequences; (D) the actual sequence alignments; and finally the parameters and other statistics gathered during the search.

Sample blastp output from comparing pir|A01243|DXCH against the SWISS-PROT database is presented below.

A. Program Introduction

The introductory output provides the program name (BLASTP in this case), the version number (1.4.6MP in this case), the date the program source code last changed substantially (June 13, 1994), the date the program was built (Sept. 22, 1994), and a description of the query sequence and database to be searched. These may all be important pieces of information if a bug is suspected or if reproducibility of results is important.

The "Searching..." indicator indicates progress that the program made in searching the database. A complete database search will yield 50 periods (.), or one period per database sequence, whichever number is smaller. When searching a database consisting of 50 sequences or more, if fewer than 50 periods are displayed and the program aborted for some reason, dividing the number of periods by 0.5 will yield the approximate percentage (0-100%) of the database that was searched before the program died. If the program had difficulty making progress through the database, one or more asterisks (*) may be interspersed between the periods at one-minute intervals.

B. Histogram of Expectations

Shown in the output below is a histogram of the lowest (most significant) Expect values obtained with each database sequence. This information is useful in determining the numbers of database sequences that achieved a particular level of statistical significance. It indicates the number of database matches that would be reportable at various settings for the expectation threshold (E parameter).

C. One-line Summaries

The one-line sequence descriptions and summaries of results are useful for identifying biologically interesting database matches and correlating this interest with the statistical significance estimates. Unless otherwise requested, the database sequences are sorted by increasing P-value (probability). Identifiers for the database sequences appear in the first column; then come brief descriptions of each sequence, which may need to be truncated in order to fit in the available space. The "High Score" column contains the score of the highest-scoring HSP found with each database sequence. The "P(N)" column contains the lowest P-value ascribed to any set of HSPs for each database sequence; and the "N" column displays the number of HSPs in the set which was ascribed the lowest P-value. The P-values are a function of N, as used in Karlin-Altschul "Sum" statistics or Poisson statistics, to treat situations where multiple HSPs are found. It should be noted that the highest-scoring HSP

whose score is reported in the "High Score" column is not necessarily a member of the set of HSPs which yields the lowest P-value; the highest-scoring HSP may be excluded from this set on the basis of consistency rules governing the grouping of HSPs (see the -consistency option). Numbers of the form "7.7e-160" are in scientific notation. In this particular example, the number being represented is 7.7 times 10 to the minus 160th power. which is astronomically close to zero.

D. Alignments

Alignments found with the BLAST algorithm are ungapped. Several statistics are used to describe each HSP: the raw alignment Score; the raw score converted to bits of information by multiplying by Lambda (see the Statistics output); the number of times one might Expect to see such a match (or a better one) merely by chance; the P-value (probability in the range 0-1) of observing such a match; the number and fraction of total residues in the HSP which are identical; the number and fraction of residues for which the alignment scores have positive values. When Sum statistics have been used to calculate the Expect and P-values, the P-value is qualified with the word "Sum" and the N parameter used in the Sum statistics is provided in parentheses to indicate the number of HSPs in the set; when Poisson statistics have been used to calculate the Expect and P-values, the P-value is qualified with the word "Poisson". Between the two lines of Query and Subject (database) sequence is a line indicating the specific residues which are identical, as well as those which are non-identical but nevertheless have positive alignment scores defined in the scoring matrix that was used (the BLOSUM62 matrix in this case). Identical letters or residues, when paired with each other, are not highlighted if their alignment score is negative or zero. Examples of this would be an X juxtaposed with an X in two amino acid sequences, or an N juxtaposed with another N in two nucleotide sequences. Such ambiguous residue-residue pairings may be uninformative and thus lend no support to the overall alignment being either real or random; however, the informativeness of these pairings is left up to the user of the BLAST programs to decide, because any values desired can be specified in a scoring matrix of the user's own making.

BLASTP 1.4.6MP [13-Jun-94] [Build 13:58:36 Sep 22 1994]

Reference: Altschul, Stephen F., Warren Gish, Webb Miller, E
and David J. Lipman (1990). Basic local alignment search to
215:403-10.

Query = pir|A01243|DXCH 232 Gene X protein - Chicken (frag
(232 letters)

Database: SWISS-PROT Release 29.0
38,303 sequences; 13,464,008 total letters.
Searching.....d

Observed Numbers of Database Sequences Satisfying
Various EXPECTation Thresholds (E parameter values)

```
      Histogram units:      = 31 Sequences      : less than

EXPECTation Threshold
(E parameter)
   |
   V   Observed Counts-->
 10000 4863 1861 |=============================================
  6310 3002  782 |=======================
  3980 2220  812 |========================
  2510 1408  303 |=========
  1580 1105  393 |============
  1000  712  179 |=====
   631  533  161 |=====
   398  372   80 |==
   251  292   73 |==
   158  219   50 |=
   100  169   32 |=
  63.1  137   18 |:
  39.8  119    9 |:
  25.1  110    6 |:
  15.8  104    9 |:
>>>>>>>>>>>>>>>>>>>  Expect = 10.0, Observed = 95  <<<<<<
  10.0   95    4 |:
  6.31   91    3 |:
  3.98   88    1 |:
  2.51   87    3 |:
  1.58   84    0 |
  1.00   84    2 |:
```

```
Sequences producing High-scoring Segment Pairs:

sp|P01013|OVAX_CHICK GENE X PROTEIN (OVALBUMIN-RELATED) (...
sp|P01014|OVAY_CHICK GENE Y PROTEIN (OVALBUMIN-RELATED).
sp|P01012|OVAL_CHICK OVALBUMIN (PLAKALBUMIN).
sp|P19104|OVAL_COTJA OVALBUMIN.
sp|P05619|ILEU_HORSE LEUKOCYTE ELASTASE INHIBITOR (LEI).
sp|P80229|ILEU_PIG   LEUKOCYTE ELASTASE INHIBITOR (LEI) (...
sp|P29508|SCCA_HUMAN SQUAMOUS CELL CARCINOMA ANTIGEN (SCC...
sp|P30740|ILEU_HUMAN LEUKOCYTE ELASTASE INHIBITOR (LEI) (...
sp|P05120|PAI2_HUMAN PLASMINOGEN ACTIVATOR INHIBITOR-2, P...
sp|P35237|PTI_HUMAN  PLACENTAL THROMBIN INHIBITOR.
sp|P29524|PAI2_RAT   PLASMINOGEN ACTIVATOR INHIBITOR-2, T...
sp|P12388|PAI2_MOUSE PLASMINOGEN ACTIVATOR INHIBITOR-2, M...
sp|P36952|MASP_HUMAN MASPIN PRECURSOR.
sp|P32261|ANT3_MOUSE ANTITHROMBIN-III PRECURSOR (ATIII).
sp|P01008|ANT3_HUMAN ANTITHROMBIN-III PRECURSOR (ATIII).

WARNING:  Descriptions of 80 database sequences were not rep
          limiting value of parameter V = 15.


  ... alignments with the top 8 database sequences deleted .

>sp|P05120|PAI2_HUMAN PLASMINOGEN ACTIVATOR INHIBITOR-2, PLA
            (MONOCYTE ARG- SERPIN).
            Length = 415
```

```
 Score = 176 (80.2 bits), Expect = 1.8e-65, Sum P(4) = 1.8e-
 Identities = 38/89 (42%), Positives = 50/89 (56%)

Query:     1 QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTK
             +I +LL   S D DT +VLVNA+YFKG WKT F  +     PF V
Sbjct:   180 KIPNLLPEGSVDGDTRMVLVNAVYFKGKWKTPFEKKLNGLYPFRVNS

Query:    61 SFNVATLPAEKMKILELPFASGDLSMLVL 89
             N+  +   K +ILELP+A      L+L
Sbjct:   240 KLNIGYIEDLKAQILELPYAGDVSMFLLL 268

 Score = 165 (75.2 bits), Expect = 1.8e-65, Sum P(4) = 1.8e-
 Identities = 33/78 (42%), Positives = 47/78 (60%)

Query:   155 ANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPE
             AN +G+S    L +S+  H A ++++E+G E A  TG +   +
Sbjct:   338 ANFSGMSERNDLFLSEVFHQAMVDVNEEGTEAAAGTGGVMTGRTGHG

Query:   215 IKHNPTNTIVYFGRYWSP 232
             I H  T  I++FGR+ SP
Sbjct:   398 IMHKITKCILFFGRFCSP 415

 Score = 144 (65.6 bits), Expect = 1.8e-65, Sum P(4) = 1.8e-
 Identities = 26/62 (41%), Positives = 41/62 (66%)

Query:    90 LPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKY
             + D  + LE +E  I ++KL +WT+ + M +  V+VY+PQ K+EE Y
Sbjct:   272 IADVSTGLELLESEITYDKLNKWTSKDKMAEDEVEVYIPQFKLEEHY

Query:   150 LF 151
              F
Sbjct:   332 AF 333

 Score = 61 (27.8 bits), Expect = 1.8e-65, Sum P(4) = 1.8e-6
 Identities = 10/17 (58%), Positives = 16/17 (94%)

Query:    81 SGDLSMLVLLPDEVSDL 97
             +GD+SM +LLPDE++D+
Sbjct:   259 AGDVSMFLLLPDEIADV 275


WARNING:  HSPs involving 86 database sequences were not repo
          limiting value of parameter B = 9.


Parameters:
  V=15
  B=9
  H=1

  -ctxfactor=1.00
  E=10

  Query                              -----  As Used  -----     -----
  Frame  MatID Matrix name           Lambda     K       H       Lamb
   +0      0   BLOSUM62              0.316   0.132   0.370     same

  Query
  Frame  MatID  Length  Eff.Length   E     S W   T  X     E2
```

```
 +0      0      232      232      10. 57 3  11 22    0.22

Statistics:
  Query          Expected          Observed          HSPs
  Frame  MatID  High Score        High Score       Reportable
   +0      0    62 (28.2 bits)  1191 (542.5 bits)     330

  Query          Neighborhd  Word       Excluded    Failed   S
  Frame  MatID    Words      Hits         Hits     Extensions E
   +0      0       4988    5661199     1146395     4504598

  Database:  SWISS-PROT Release 29.0
    Release date:  June 1994
    Posted date:  1:29 PM EDT Jul 28, 1994
  # of letters in database:  13,464,008
  # of sequences in database:  38,303
  # of database sequences satisfying E:  95
  No. of states in DFA:  561 (55 KB)
  Total size of DFA:  110 KB (128 KB)
  Time to generate neighborhood:  0.03u 0.01s 0.04t  Real: 0
  No. of processors used:  8
  Time to search database:  32.27u 0.78s 33.05t  Real: 00:00
  Total cpu time:  32.33u 0.91s 33.24t  Real: 00:00:05

WARNINGS ISSUED:  2
```

COPYRIGHT

REFERENCES

Altschul, Stephen F. (1991). Amino acid substitution matrices from an information theoretic perspective. J. Mol. Biol. 219:555-65.

Altschul, S. F. (1993). A protein alignment scoring system sensitive at all evolutionary distances. J. Mol. Evol. 36:290-300.

Altschul, S. F., M. S. Boguski, W. Gish and J. C. Wootton (1994). Issues in searching molecular sequence databases. Nature Genetics 6:119-129.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman (1990). Basic local alignment search tool. J. Mol. Biol. 215:403-10.

Claverie, J.-M. and D. J. States (1993). Information enhancement methods for large scale sequence analysis. Computers in Chemistry 17:191-201.

Gish, W. and D. J. States (1993). Identification of protein coding regions by database similarity search. Nature Genetics 3:266-72.

Henikoff, Steven and Jorga G. Henikoff (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89:10915-19.

Karlin, Samuel and Stephen F. Altschul (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA 87:2264-68.

Karlin, Samuel and Stephen F. Altschul (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. Proc. Natl. Acad. Sci. USA 90:5873-7.

States, D. J. and W. Gish (1994). Combined use of sequence similarity and codon bias for coding region identification. J. Comput. Biol. 1:39-50.

States, D. J., W. Gish and S. F. Altschul (1991). Improved sensitivity of nucleic acid database similarity searches using application specific scoring matrices. Methods: A companion to Methods in Enzymology 3:66-70.

Wootton, J. C. and S. Federhen (1993). Statistics of local complexity in amino acid sequences and sequence databases. Computers in Chemistry 17:149-163.
